Week 8 Report

# Nested Chinese restaurant process(nCRP) and hierarchical LDA (hLDA)

Authors: David M. Blei, Thomas L. Griffiths…

Link: <https://proceedings.neurips.cc/paper_files/paper/2003/file/7b41bfa5085806dfa24b8c9de0ce567f-Paper.pdf>

The **nCRP** provides a nonparametric prior for the **hLDA** model, replacing the standard Dirichlet distribution used in traditional LDA. This approach allows for a nonparametric model that can organize topics into a hierarchical structure.

## 1. Generative Process

### 1.1 Model Parameters

#### Hyperparameters:

*γ*: Controls how likely a customer will sit at a new table

*α*: Dirichlet distribution parameter. Used to sample *θ*

*η*: Dirichlet distribution parameter. Used to sample

#### Variable:

*θ*: Document-specific topic distribution

*β*: Topic-specific word distribution

*z*: The topic assignment for each word in the document. The topic is drawn from the multinomial distribution defined by the document’s topic proportions *θ*

*c*: The path through the hierarchy that a document takes.

*N*: The number of words in the document.

*W*: The size of the vocabulary.

### 1.2 Assumptions

#### 1.2.1 Bag of words assumption

1.2.2 Exchangeability of Documents: Order of document does not matter

1.2.3 Fixed sized tree: The hierarchy is restricted to a fix length of L. However, this can be relaxed.

### Outline of Algorithm

Figure (1): (a) Represents the finite tree derived from the nCRP. (b) Graphical representation of the model hLDA.

|  |
| --- |
| **hLDA outline:** |

For each document:

* 1. Start at the root of the topic tree
  2. Traverse the hierarchy level by level:
     1. At each level, select a table using the ordinary CRP
        1. Choose an existing table (previously visited topic) with probability proportional to the number of documents already seated there.
        2. Choose a new table (introduce a new topic) with probability proportional to *γ*.
  3. Once the path is chosen, sample the topic proportions from a Dirichlet distribution .
  4. For each word in the document:
     1. Sample a topic *z* from the multinomial distribution over topics , based on the chosen path.
     2. Sample the word from the multinomial word distribution associated with the topic *z*.

## 2. Inference Process

The inference processes makes use of Gibbs sampling to accesses the 2 variables of interest:

: Topic assignment of the *n*th word in the *m*th document to one of the *L* available topics.

: Restaurant corresponding to the *l*th topic in document *m*.

Conceptually, the Gibbs sampler is split into 2 parts. First, given the current state of the CRP, we sample of the underlying LDA model. Secondly, given the values of the LDA hidden variables, we sample with the variables which are associated with the CRP prior.

### 2.1 Sampling level allocations,

### 2.2 Sampling path,

Note that the *cm* must be drawn as a block, meaning the entire path for document *m* is sampled together.

**Posterior used:**



Figure 2: Posterior

A math equations with black text

Description automatically generated with medium confidence